

DOCUMENT RESUME

ED 188 701

JC 800 394

AUTHOR McCullough, K. Owen, Ed.
TITLE Proceedings: Conference on the Community/Junior College (University of Tennessee, Knoxville, Tennessee, April 25-26, 1975).
INSTITUTION Tennessee Univ., Knoxville. Coll. of Education.
PUB DATE 26 Apr 75
NOTE 59p.
AVAILABLE FROM Bureau of Educational Research, University of Tennessee, Knoxville, TN 37916 (\$2.00)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Community Colleges: Criterion Referenced Tests; *Evaluation: *Evaluation Criteria: *Evaluation Methods: Evaluators: Information Utilization: Measurement Objectives: Models: Nontraditional Education: Norm Referenced Tests: Program Evaluation: Test Bias: *Testing
IDENTIFIERS Domain Referenced Tests

ABSTRACT

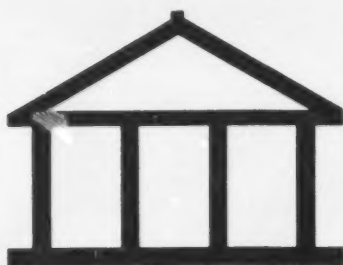
The 1975 Community/Junior College Conference sought to define and describe evaluation, discuss measuring devices, and underscore the ethical reasons for fair evaluation. William P. Gorth began by discussing the purpose of evaluation--to provide data to a specified decision-maker for decision-making--and the implications which could be drawn from that basic purpose. He enumerated the 12 steps in an evaluation cycle and provided a hypothetical example. He considered the roles of the evaluator and the decision-maker, contrasted generalized and specialized evaluation, and established what evaluation is not. Next, Peter E. Schriber discussed testing and its place in the process of evaluation. He concentrated on norm-referenced, criterion-referenced, and domain-referenced testing and their different uses, construction, and relation to the content domain. Eva I. Baker placed evaluation in the context of higher education and presented specific means of improving it to provide better and fairer evaluation data. Finally, Richard G. Allan addressed the issue of evaluating non-traditional programs. He described several hypothetical programs, alternatives for program evaluation, and criteria for determining which evaluation method is most appropriate to the particular program. The conference proceedings include a roster of participants. (AYC)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED188701

PROCEEDINGS CONFERENCE ON THE COMMUNITY/JUNIOR COLLEGE

APRIL 25, 26, 1975 THE UNIVERSITY OF TENNESSEE



"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Ken McCullough

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT
OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

JC 800 394

Contents

Preface /

K. Owen McCullough / 4

1975 Community / Junior College

Conference Report /

Earl M. Ramer / 5

Evaluation: What Is It,

How Is It done? /

William Phillip Gorth / 8

Testing: A Tool For Evaluation /

Peter E. Schriber / 21

Fair Measurement

and Higher Education /

Eva L. Baker / 33

Methods of Evaluation In

Non-Traditional Settings /

Richard G. Allan / 45

Conference Roster / 55

Preface

K. OWEN MCCULLOUGH

Editor and Planning Committee Chairman

University of Tennessee, Knoxville

June, 1975

It seems that there are always a couple of ideas or concepts which occupy center stage in the continuing drama of education. Teacher tenure, collective bargaining, learning modules, self-instructional packages, multi-media techniques are merely a few of the items which have continually interested us as professional educators. One which seems to be most pervasive, however, one which seems always with us, is *evaluation*. Regardless of the level, whether it is at the discussion level or the procedural level, whether we are concerned with programs or individual students, the question of "how to evaluate?" is with us constantly.

In the following presentations, you will see, then, that our '75 Conference was primarily concerned with the *how* part of evaluation. We defined it, described it, talked about measuring devices, and even touched on the ethical reasons for fair measurement, but our concern was always pragmatic, utilitarian in nature. We were seeking answers to specific questions; we wanted clarity; we wanted release from the confusion which has surrounded evaluation methodology for too long.

Because of this, our speakers faced a difficult audience, polite and appreciative, but difficult and impatient with generalities. Many questions were asked; most were answered. But there were a few that seemed unanswerable, or a few that had multiple answers, and there were, as always, many that remained unasked. It would have been helpful, perhaps, had the speaker-participant interplay been recorded for these proceedings.

The following presentations are offered more as a chronicle of our conference than as an evaluation handbook. Our reach, it seems, was farther than our grasp. We intended to sneak up and snare the topic as we might a rabbit. But we only touched it, and, like a fox, the essence of evaluation remains to be caught by only the wise and most persistent hunters.

1975

Community/Junior College Conference Report

EARL RAMER

For the 1974-75 Planning Committee

Consistent with Planning Committee policy to meet on the UTK campus as well as on the campuses of other state institutions, especially community colleges, the 1975 Conference convened on the UTK campus on April 25-26, 1975. Continuing the evaluation theme from the preceding year, but attempting several specific applications, the conference attracted national leaders as speakers and consultants, and good attendance from the state and region.

As in former years, the 1975 *Proceedings* include major conference presentations. At the general assembly Friday afternoon, William Gorth, President of the National Evaluation Systems, Inc., spoke on the topic, "Evaluation of Student Learning: What it is, how it is done." That evening, Peter Schriber, of the same agency, discussed "Testing: One evaluation strategy."

At the opening session Saturday morning, Eva Baker, UCLA, spoke on the topic "Construction and Use of Evaluation Instruments." At the concluding session Saturday afternoon, Richard Allan, Vice-president, National Evaluation Systems, Inc., addressed the conference on the topic, "Methods of Evaluation in Non-traditional Settings."

Much credit for whatever success the conference enjoyed must go to members of the Reactions Panels. Membership on these panels included James Amburgey, Hiwassee College; Uriah Clemmer, Columbia State Community College; Jane Dawson, Martin College; Jeanette Held, Martin College; Bob McElroy, Walters State Community College; Arthur McFadden, Chattanooga State Technical Community College; Anne Minter, Roane State Community College; Lewis Moore, Columbia State Community College; James Perry, Roane State Community College; Paul Ritch, Chattanooga State Technical Community College; Darrell Simmons, Walters State Community College; and Sister Henry Suso, Aquinas Junior College.

Ralph Norman, Associate Vice-Chancellor for Academic Affairs, UTK welcomed the Conference to the UTK campus. Presiding at the major sessions and introducing speakers were David F. Adkisson, President, Cleveland State Community College; Charles Branch,

President, Chattanooga State Technical Community College; Wray Buchanan, Vice-Chancellor for Academic Affairs, Tennessee State Board of Regents; Jack Campbell, President, Walters State Community College; Cuyler Dunbar, President, Roane State Community College; James McComas, Dean, College of Education, UTK; Alvin Nielsen, Dean, College of Liberal Arts, UTK; John Prados, Vice President for Academic Affairs, UTK; and Tom Stovall, Associate Director of Academic Affairs, Tennessee Higher Education Commission.

The Planning Committee, identified at another point in the *Proceedings*, joins sponsoring agencies and institutions in expressing appreciation to all of these individuals and groups identified above. Thanks must go, of course, to Ken McCullough, Conference Director, who served faithfully and efficiently in carrying out all of those numerous responsibilities necessary to the success of any conference of this type. As in former years, Dean Jim McComas and Associate Dean Bill Coffield worked with other University officials in making financial assistance available. For this we are grateful.

For the fifth year, now, the UTK Bureau of Educational Research and Service has assumed responsibility for the publication of these *Proceedings*. Dr. Charles Peccolo and Bureau Staff Members are to be commended for their assistance with this task. Special thanks must be given to Barbara Arwood of the Bureau who, from the beginning, has related the *Proceedings* format meaningfully to conference purpose and content.

Again, this year, participation from public and private two- and four-year campuses throughout the state has been excellent. The Planning Committee is gratified with the extent of this now customary participation and desires and seeks it for the 1976 Conference, planning for which is already underway. If you have suggestions of any kind for the next Conference, please communicate with Ken McCullough.

Single copies of the 1975 *Proceedings* are being mailed to all Conference participants. Other copies at \$2.00 each may be ordered from UTK Bureau of Educational Research.

CONFERENCE PLANNING COMMITTEE:

Dave Adkisson, President
Cleveland State Community College
Cleveland, Tennessee

Cuyler Dunbar, President
Roane State Community College
Harriman, Tennessee

LeRoy Graf, Professor & Head, History
1101 McClung Tower
The University of Tennessee, Knoxville

Roger Haskell, Professor, Vocational-Technical Education
113 Henson Hall
The University of Tennessee, Knoxville

Bill Hodges, Faculty Member
Walters State Community College
Morristown, Tennessee

John Lovell, Professor, Education Administration
and Supervision
204 Henson Hall
The University of Tennessee, Knoxville

K. Owen McCullough, Professor
Continuing and Higher Education
15 Henson Hall, The University of Tennessee, Knoxville

John McDow, Dean, Admissions and Records
301 Student Services Building
The University of Tennessee, Knoxville

Ohmer Milton, Director,
Learning Research Center
The University of Tennessee, Knoxville

Earl Ramer, Professor
Continuing and Higher Education
15 Henson Hall, The University of Tennessee, Knoxville

Tom Ryan, Professor, English Education
216A Claxton Education Building
The University of Tennessee, Knoxville

Evaluation: What Is It, How Is It Done?

WILLIAM PHILLIP GORTI

President

National Evaluation Systems, Inc.

The word evaluation strikes fear and confusion in the hearts of many educators, primarily because it is a concept often misunderstood. Yet it is in great demand both by educators and taxpayers. The confusion about education evaluation is, in part, a result of the short period in which it has been a discipline in its own right. Its history extends back to about the early 1960's. Only as recently as 1970 was a specific division of the American Educational Research Association formed which focused on the topic of evaluation.

Therefore, an introduction to education evaluation should attempt to dispel the confusion by articulating a fundamental vocabulary and explaining some underlying concepts related to education evaluation. As in any new discipline, the vocabulary used by its different practitioners varies widely. Similar concepts are given different names and different concepts the same name by evaluators. What is needed is a straight-forward vocabulary which can be used to discuss basic concepts in education evaluation. The basic concepts must be described in familiar terms to allow persons interested in education evaluation to share and discuss ideas.

The basic purpose of this paper is to present the vocabulary and concepts necessary for an adequate discussion of education evaluation. The paper will begin by discussing the basic purpose of education evaluation. It will present the purpose of evaluation in an historical perspective and will enumerate briefly its implications. The balance of the paper will be concerned with a description of a basic set of concepts related to the process of education evaluation—concepts which would be part of any evaluation. Last, several controversial issues related to education evaluation will be discussed from the point of view of the practitioner and the potential consumer of evaluation services.

Purpose of Evaluation*

Although agreement in the purpose of education evaluation does not exist, a widely accepted purpose may be stated as follows:

The discussion of educational evaluation which is used in this paper is closely related to the work of Thomas E. Hutchinson at the University of Massachusetts.

TO PROVIDE DATA TO A SPECIFIED DECISION MAKER FOR DECISION MAKING. This is the fundamental purpose for education evaluation and clearly conveys the intent of many practitioners in their use of the term.

Insight into the stated purpose of education evaluation can be gained by examining each of its parts. First, the words "to provide" imply that something will be supplied which may be of use to the education practitioner. "To provide" need not be limited to the familiar notion of an evaluation report. Meetings, presentations, media displays, and other strategies for making available the result of the evaluation are all acceptable under the statement of purpose.

Second, the word "data" is the fundamental commodity being provided as a result of education evaluation. "Data" are not limited to specific types of data. The data may include any type of information. It is only the specific situation which determines which type of data is supplied. The data may include achievement and aptitude tests, inventories of attitudes, questionnaires about demographic information, interviews about personal opinions, or physical measurements.

Third, the words "specified decision maker" are central to much of the discussion of education evaluation which will follow. Data should not be gathered haphazardly without a specified decision maker in mind. The decision maker may be either an individual, such as a teacher, or a group of persons, such as a Board of Regents or a Board of Trustees. The decision maker must be specified in advance of the evaluation activities and must be carefully considered at each step of the evaluation process.

Fourth, and last, "for decision making" implies that the decision maker, either an individual or a group of persons, will use the data provided as a basis for choosing among alternative courses of action. Data from the evaluation should be the bases upon which the decision maker is able to choose the most different alternative courses of action.

Historically, as education evaluation has matured, its purpose has changed considerably. During the early portion of this century, education evaluation was equated with the concept of testing or measurement. During this period many authors referred to the development of a test and its use as the definition of education evaluation. During the 1940s and 1950s the definition of evaluation was expanded by authors such as Ralph Tyler, whose work focused on curriculum development. Tyler defined evaluation as follows: "The process of evaluation is essentially the process of determining to what extent educational objectives are actually being realized by the program of curriculum and instruction."

In the early 1960s and 1970s, another perspective on education evaluation developed. Authors such as Robert Stake advocated the view that education evaluation should be "the process of providing the professional judgment on educational activities." The professional judgment came from reviewing the performance of an educational program and judging its value or acceptability.

Since the early 1960s, the definition of evaluation, on which this paper will focus, has been developed and widely used (Provus, Stufflebeam, Hutchinson, Gorth). It is important to note that the earlier and current alternative approaches to defining education evaluation are not identical with the purpose described in this paper, which is, "to provide data to a specific decision maker for decision making." This is broader than either testing or comparing performance with educational objectives, and it is different from making professional decisions.

Implications of the Purpose

Important implications can be drawn from the basic purpose specified for education evaluation. These implications must be kept in mind during the later discussions of the process of evaluation and during the application of evaluation to a specific situation. First, and by far the most important implication of the purpose of education evaluation, is that the data which are collected are useful to a specific decision maker. The decision maker should understand these data, find them credible, and use them. If he does, the evaluation is said to have *decision maker validity*. Education evaluation must attempt to ensure in every way possible that the data can be used by the decision maker for decision making. Providing data for hypothetical decision makers or general decision-making audiences is a weak substitute for identifying specific decision makers and tailoring the data provided to each of them.

Second, the data provided by education evaluation ought to be associated with specific decisions which have been identified by the decision maker. If data are collected which are not applicable, there is a high probability that those data will not be used for decision making. So it is obvious that in designing the process for providing data, some provision should be made for ensuring that data are related to anticipated decisions. This is frequently not the case in evaluation. Often collected data are not useful to the specified decision maker or to any decision maker. To the extent that data provided are not used, the portion of the resources allocated to provide those data are wasted; therefore, the evaluation is inefficient. The degree to which data, provided by the evaluation, are used by the decision maker determines the *efficiency* of the evaluation.

Third, although an education evaluation may provide data which are related to decisions made by the decision maker, it is further important that these data are directed at important decisions to be made rather than relatively unimportant decisions. An evaluation which is able to provide data for an important decision is a more successful evaluation than one which is able to provide data for only an unimportant decision. The degree to which data are provided for important decisions rather than unimportant decisions is a measure of the *focus* of an education evaluation.

Fourth, an evaluation should attempt to be as extensive as possible in providing all data necessary for all decisions made by the decision maker. Although education evaluations strive for this goal, it is obvious that the practical constraints caused by the amount of resources for the evaluation are going to limit in some way the amount of data provided. However, the degree to which data are provided for the total set of decisions made by the decision maker is the measure of the *completeness* of an evaluation for that specific decision maker.

Fifth, if the education evaluation is being designed to provide data for a specific decision made by a specific decision maker, it is clear that those data may not be useful to other decision makers. This situation should not worry evaluators. An evaluation may not provide useful data to all potential decision makers because each decision maker may make different kinds of decisions. A first-year teacher within a college will have to make different decisions about a course than the president of the college. Therefore, an evaluation designed solely for the president of the college may not provide information that is relevant to the individual teacher. This is not a shortcoming of the evaluation process, but just a specific observation that data designed to be of maximum usefulness to one decision maker may not be useful to all decision makers.

Sixth, throughout the discussion of education evaluation, there will be a concern for maintaining a high degree of quality of the data provided. Traditional concepts of the *validity* and *reliability* of measurements must be of concern to the evaluator. Validity is the degree to which the measurement actually relates to the phenomenon that it is designed to measure. Reliability is concerned with the consistency of the measurement from one time to another. In evaluation a high premium is put on the validity of data so that they are clearly related to the aspects of the educational activity for which they are designed.

The Steps of an Evaluation Cycle

The process of evaluation will be described in general steps

which form the evaluation cycle. These major steps are conceptualized as twelve major concepts or sets of activities. They are displayed in Figure 1. Each step has a specific purpose and is carried out by specific persons associated with the evaluation. This section of the paper will list each of the twelve steps and give a short description of the persons involved as well as the basic activities associated with each of the steps. More detailed sub-steps, which make up each of these major steps, are available but will not be presented here.

Step 1: The *enterprise* is the name given to the educational activity, such as a course, a program, a college department, or an administrative unit, about which the decision maker will make decisions and about which the data will be collected. Therefore, the first step in any education evaluation must be the definition of the educational enterprise in terms of a short statement or paragraph. The person initiating the evaluation provides the statement to the evaluator. The statement of the enterprise may take the form of a short description of what the enterprise is. It may also be useful to identify the things that are not included in the enterprise and that are not to be considered by the evaluation.

Step 2: The *resources* available for the evaluation must be determined in detail by the initiating person. The evaluator does not provide resources. The resources must be obtained from the agency purchasing the evaluation services. The agency must decide what resources the evaluator will have at his disposal. These resources may include both funds which would be under the fiscal control of the evaluator and resources of the agency such as personnel, duplication facilities, and other resources (which would be available in specific amounts to the evaluator for the evaluation activities).

Step 3: The *decision maker*, who makes decisions about the enterprise and to whom data are to be reported, is identified by the person initiating the evaluation. The evaluator is not responsible for choosing the person or the group to whom data are to be provided. The evaluator provides a structured procedure by which the person initiating the evaluation can carefully consider each of the potential decision makers and systematically select a decision maker for whom data will be collected. The decision maker, as discussed in this paper, may be an individual or a group of individuals acting together. The initiator may or may not select himself to be a decision maker to whom data are reported.

Step 4: The *goals* of the enterprise are elicited from each decision maker by the evaluator. A goal is a general statement of what is to be accomplished. Goals are not stated by the evaluator. A process exists by which the evaluator attempts to ensure that the decision maker is thorough in articulating goals and identifies the relative

Evaluation

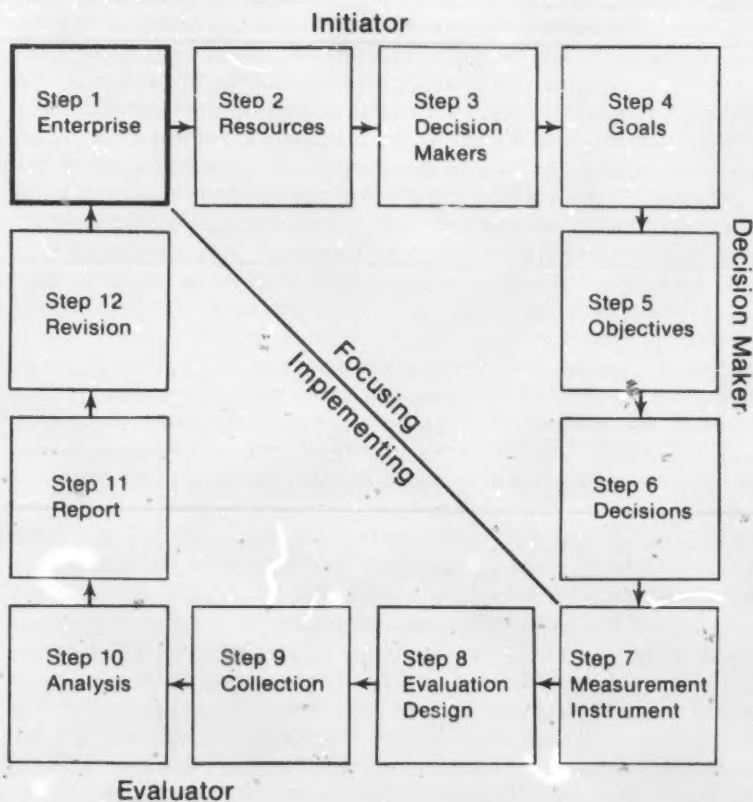


Figure 1

importance of each. Having identified the goals and ranked them from more important to less important, the evaluator is in a position to focus the data of the evaluation on important goals for which data are not available.

Step 5: The *objectives* associated with each goal are elicited from the decision maker by the evaluator. An objective is a specific, operational statement of what is to be accomplished or attained. The decision maker must enumerate what objectives are related to each of the important goals that he has articulated, thereby creating an operational and observable description of goals in terms of the objectives. These objectives are not stated by the evaluator, but the evaluator does lead the decision maker through a step-by-step process to ensure that the goals are thoroughly and completely articulated in terms of operational characteristics which represent the true feelings of the decision maker. The objectives are then ranked in order of importance.

Step 6: The *decisions* to be made about each objective or set of objectives are elicited from the decision maker by the evaluator. The decision maker must anticipate what kinds of decisions he plans to make in relation to each of the specific objectives and goals which he has articulated. The evaluator guides the decision maker through a process of identifying important decisions related to important objectives so that data gathering can focus on those decisions which are of most importance and for which no other sources of data would be available except the evaluation.

Step 7: The *measurement technique*, i.e., data gathering instrument, is developed by the evaluator to successfully measure the individual objective which has been articulated by the decision maker. Each of the measurement instruments is carefully tied to one of the objectives previously defined by the decision maker. It is the evaluator's expertise in developing measurement instruments which is used at this point to create a measurement technique which will successfully measure that which would be useful to the decision maker.

Step 8: The *design* of the measurement instrument to collect data is the evaluator's responsibility. The design is concerned with collecting information which will be useful to the decision maker for the specific decisions which he has identified. It is important that the evaluation design provide data on a timely basis to the decision maker for each specified decision. Further, the decision maker should have an opportunity to review the design and understand its structure so that he can approve or disapprove of its appropriateness for gathering the data which he feels are appropriate.

Step 9: The *collection* of data is carried out in a systematic way by

the evaluator who is primarily responsible for this step of the evaluation. Data collection should have as little interference with the educational enterprise as possible. Further, it should be done at a minimum cost as quickly as possible.

Step 10: The analysis of the data should be performed quickly and efficiently by the evaluator. The evaluator chooses analysis procedures which are compatible with the types of data collected. However, he also checks with the decision maker to ensure that the analysis will produce information that is understandable and useful to him. Too many education evaluations report analyses which are so complex statistically that a decision maker is not able to use the data. Often, counts of the number of occurrences in behavior or percentage of behavior by different groups is the most valuable, understandable, and useful information that an evaluator provides to decision makers. The desire to provide more complex statistical analyses interferes seriously with the usefulness and the utilization of data by a decision maker because he is unsure of their meaning and implication. Although the evaluator carries out the analysis, and uses his technical expertise to be sure that it is done accurately and completely, the evaluator checks with the decision maker to ensure that an appropriate and understandable analysis has been selected.

Step 11: The report of the data to the specific decision maker is developed and written by the evaluator. This report contains detailed explanations of the relationship of all data to the decision makers' goals, objectives, and anticipated decisions so that it is clear and unambiguous to what decision the decision maker should apply the data. The report of the data therefore is a result of the efforts of the evaluator to accurately and completely portray the data which he has collected. However, the evaluator is not responsible for making decisions or suggesting alternatives for decisions. Therefore, he does not include recommendations and conclusions in the report of the data to the evaluator. The responsibility for decision making remains with the decision maker.

Step 12: The revision of the evaluation should be considered jointly by the decision maker and the evaluator after the report of the data has been made. The revision may include minor or major modifications at any one of the preceding 11 steps in the evaluation. These revisions may be requested specifically by the decision maker or may be suggested because of lack of usefulness of the data for decision making as identified by the evaluator.

A Hypothetical Example

A hypothetical example will be used to briefly illustrate the evaluation cycle. A college president is enthusiastic about the

mental health technology program at his college. He thinks the program is valuable because of the current personnel need in local hospitals and mental health clinics. He is concerned with improving the program as much as possible. Therefore, he contacts an evaluator and begins the process of evaluation.

During their first meeting the evaluator and the president complete steps 1, 2, and 3. The president of the college is concerned with providing data with the mental health program as the enterprise to be evaluated (Step 1). The mental health program does not include any of the other curriculum areas relating to human services, but will focus on the courses and activities relating to mental health technology. The president is willing to allocate \$2,000 to hire an outside evaluator and is willing to allocate approximately 5,000 photocopies, unlimited offset press copies, and a half-time secretary as resources for the educational evaluation (Step 2). The president thinks about a large number of persons associated with mental health technology program including students, faculty, department chairmen, central administrators, college trustees and taxpayers, all of whom are potential decision makers. After seriously considering all of these alternative potential decision makers, the president selects the director of the mental health technology program as the decision maker for whom the evaluation will be designed and to whom evaluation reports will be made (Step 3).

After a preliminary informal meeting between the president, program director, and evaluator, the evaluator and the program director have a working meeting. According to a specified set of evaluation procedures, the program director generates a clear, up-to-date statement of the goals which he holds for the mental health technology program and organizes them from most to least important (Step 4). Each of the *important* goals is specified in clearly defined, operational objectives (Step 5), and the decisions related to each of the objectives is specified (Step 6). The program director is responsible for articulating the goals and objectives which he sees for the program. The evaluator guides the director through a series of steps to ensure that this is carried out thoroughly. The decisions of primary concern to the program director related to the success of the curriculum in teaching the objectives that have been defined for the program. As a specific example, one important goal is: "Students will demonstrate supportive interpersonal relationships with patients." The program director has identified several specific objectives related to this goal including: "The student will establish rapport with a client by presenting a specific purpose for any meeting or interaction." The program director is concerned with establishing whether students are reaching this particular objective.

Having identified the decisions, the evaluator develops a measurement technique to measure the extent to which the mental health technology students establish rapport with a client at the beginning of any interaction (Step 7). This measurement technique might include observations of the student in an internship situation with a client or a role-playing situation in the classroom. The evaluator proposes both, along with their respective costs, to the decision maker who selects role playing. The design of the evaluation is then developed by the evaluator (Step 8).

The evaluator collects the data (Step 9), analyzes the data (Step 10), and presents a report to the program director by the end of the first semester (Step 11). The program director finds that this particular skill has been successfully taught to the students in the program. He suggests a revision of this aspect of future evaluation activities, excluding this particular decision and focusing on other decisions for which no data are available (Step 12). Therefore, the revision function has been satisfied based upon the successful provision of data to a specific decision maker for a specified decision. After all of the decisions have been made, a revision of the total evaluation for the future is made so that the evaluation provides data for other decisions.

Distinction in Roles

Much of the literature about evaluation is confusing to a reader because a clear distinction is not made between the role and responsibilities of the evaluator versus the role and responsibilities of the decision maker. The educational decision maker is the individual or group that uses the data collected by the evaluation. These data are used as bases for making decisions. These data are normally collected in such a way that they are related to specific decisions which the decision maker must make about an educational enterprise. The educational decision maker makes decisions or judgments based upon his own personal value system, when he chooses among the alternatives available to him. The decision maker, therefore, has the responsibility to use the data to help him accomplish his overall goals and objectives for the enterprise in terms of his own personal frame of reference. The responsibility for decision making clearly rests with the decision maker.

The education evaluator carries out the tasks defined in the evaluation cycle. He is responsible for coordinating activities, developing materials, and supervising the overall evaluation activities. He is providing data to someone else, namely the decision maker.

In this paper the evaluator and the decision maker are two

separate individuals. In thinking about education evaluation, it is important clearly to separate the role of the evaluator and the decision maker so that the responsibilities and activities associated with each role can be clearly described and discussed.

The discussion of the evaluation cycle emphasized the situation in which decision-maker roles and evaluator roles are clearly the responsibilities of two separate individuals. However, in the real world, there is often an overlap between the two roles. This overlap causes substantial confusion in the discussion of evaluation. Clarification of two frequently occurring types of overlap is in order to help clarify the roles and responsibilities of education evaluators and education decision makers. Type 1 overlap occurs in educational agencies which employ evaluators or in funded projects which hire their own evaluator. In both situations the evaluator is often either a colleague or a subordinate of the persons who have hired him and are responsible to some degree for his activities. Therefore, the evaluator is in a position of being potentially influenced by other members of the organization to which he belongs. He may be asked to carry out evaluation activities which are different from the best possible evaluation procedures. These situations create difficulties for evaluators in their efforts to present accurate and unbiased information to a specified decision maker.

Type 2 overlap is a more extreme case of overlap between the evaluator and the decision maker. This overlap occurs in every classroom in the country. A classroom teacher has responsibilities of an evaluator and of a decision maker for his students. Therefore, a teacher must serve both roles, i.e., evaluator and decision maker. Because a teacher must play both roles, there must be a clear understanding by the teacher of the evaluation methods so that his role as evaluator is carried out successfully. At the same time the teacher uses the data provided by the evaluation for decision making in the classroom. Clearly, there are different roles associated with evaluating and decision making. The teacher must be sensitive to this difference in order to be successful in the performance of each.

Generalized Versus Specialized Evaluation

Confusion in the literature of education evaluation also results from the lack of clarity as to the purpose to which a particular approach of evaluation is designed to satisfy. The purpose for the evaluation approach described in this paper is generalized. There have been no specific restrictions stated in the purpose for the evaluation method which limits the type of enterprise that will be considered nor the type of data which will be provided nor the type of decision maker to whom data will be provided.

However, a large number of evaluation methods have been developed to handle only specific situations in education. These evaluation methods are focused on a specific enterprise such as the education of the handicapped or guidance procedures. They may focus on a particular type of decision maker, such as an administrator rather than a teacher or a student. They may be concerned with providing only a specific set of data, such as achievement test data or informal observational data. Each of these specialized evaluation methods may provide a successful model for the detailed evaluation of a specific aspect of an educational enterprise or a specific educational decision.

It is important to separate the specialized evaluation methods from the generalized method. Specialized evaluation methods are not appropriate to all situations and must be applied only in situations for which they are designed. Confusion exists when specialized evaluation methods for different enterprises or different types of decision makers are compared, and their specializations are not taken into account.

What Evaluation Is Not

Confusion in the discussion of evaluation sometimes is created by thinking of other activities as evaluation activities when they are not. Education evaluation has a very clearly stated purpose: "to provide data to specified decision makers for decision making." Other activities associated with education have different purposes, although they may have some activities which overlap.

In particular, curriculum development often is confused with evaluation. Curriculum development is concerned with the development of materials and procedures to facilitate student learning. In the process of developing these materials and procedures, the curriculum developer often creates objectives and in some cases creates measurement instruments. They are similar to the objectives and measurement instruments which would be developed by the evaluator. However, the curriculum developer is concerned primarily with the development of the content to teach the objectives and with methods by which to present the content. Therefore, curriculum development is not synonymous with education evaluation.

Similarly education planning is concerned primarily with identifying or creating procedures for alleviating deficiencies identified in some educational enterprises. Objectives and goals may be written as a way of articulating particular deficiencies. However, planning is different from evaluation.

Often the notion of public relations is confused with evaluations.

Public relations is concerned with presenting a positive and favorable image about an enterprise to some audience. Public relations is not concerned with presenting a totally balanced picture of an educational enterprise. Public relations also uses a variety of techniques including kinds of media presentations and narrative descriptions which may be substantially different from those used in education evaluation. (Evaluators usually do not make very good public relations persons because their primary concern is with presenting balanced empirical data for specific decision makers, whereas public relations is concerned with presenting positive and persuasive information to a general audience.)

Summary

Education evaluation is a new area of disciplined inquiry which is slowly developing a vocabulary and a set of basic concepts which are becoming more widely accepted in the profession. Evaluation has a specified purpose which is generally accepted as: "to provide data for a specified decision maker for decision making." From this purpose a series of activities and steps in an evaluation cycle can be generated which describe major conceptual components associated with evaluation. These components can be given relatively straightforward names and can be discussed as a logical sequence which contains prerequisite activities and a systematic sub-structure. This series of steps has been presented at a general level.

The evaluation method described in this paper clearly differentiates between the role of an evaluator and the role of decision maker. It also points out that education evaluation is different from other activities in education such as curriculum development, planning, and public relations. The confusion associated with the discussion of evaluation results from an inconsistent use of its concepts. By clearly saying what evaluation is, it is hoped that persons discussing evaluation activities can clarify their use of the term and more successfully carry out these activities.

Testing: A Tool For Evaluation

PETER E. SCHRIBER
*Vice-President for
Educational Services
National Evaluation Systems, Inc.*

Testing is a word with few positive connotations to many educators. Many of us tend to feel that we are imposing a constraint on our teaching and learning situations because we "resort" to the use of tests. We often refer to tests disdainfully as the "necessary evil." We characterize our own instructor-made tests as those things whose chief value is to make our students bear down and study and as a chief means to justify the grades we assign. Sometimes we get some individual diagnosis out of the test results, but often we get no more than this.

We take a narrow view of tests and often consider them outside the instructional process. Tests we construct ourselves or even buy commercially are often used for purposes inappropriate to the test design and content. We are unprepared to use the tests effectively for all the useful purposes to which tests can be put for instructional decision making.

My point is that all too often we have only a surface knowledge of a phenomenon and conceptualize it in a very traditional way. Socrates is credited with saying that the unexamined life is not worth living. We may paraphrase that and say that the test which is not well planned and designed for a particular purpose or which is not used for the purpose for which it was constructed is not worth giving.

Testing is an integral part of the instructional process. It is an important part of instructional management. Generally, however, we treat testing as an imperfected methodology. We recognize that testing is, at best, a simulation of real world activities and, as such, is an indicator of learning but not a complete measure of learning. Coupling this fact with misuse and abuse of tests, we are likely to regard tests as a necessary evil, something a little bit foreign to the main stream of instruction and learning.

You have all heard the following things said about tests and many of us have repeated them:

"You should not teach for the tests."

"Tests should provide a spread of scores so that one can assign grades using tests as a basis."

"A good test should challenge my best students."

"The most important information from a test is what it says about the individual."

All these statements are false. They are also all true! This paradox is resolved when we point out that there are two major types of testing—one is norm-referenced testing which we call NRT for short and the other is criterion-referenced testing known as CRT for short. Now don't get too attached to the name CRT. It has been around for about 12 years and is now giving way to a new term—domain-referenced testing or DRT—we'll get to that term shortly.

One might begin by pointing out that NRT tests are those designed to differentiate among individuals on a particular attribute which the test is designed to measure. The tests may be measures of actual achievement or measures of potential achievement which is often referred to as ability or aptitude. The attribute is usually very broad, such as reading ability, chemistry achievement, general academic aptitude, mathematics achievement, mathematics aptitude, and so on. The results of such tests are rankings, and we speak of an individual or group having a particular percentile rank. Thus, we might say the following things about NRT tests:

"The tests should provide a spread of scores so that I can rank the students and select the top 25% for admission to our new special physical science program."

"A good test should challenge even my best students."

"My class is below average for a first year biology class, based on all biology students in the Mid West."

Before I contrast NRT with DRT, let's take a closer look at how testing fits into evaluation. Dr. Gorth has identified and enumerated 12 steps in a systematic evaluation. A test is a tool for evaluation. It is the focus of Step 7, the Measurement Technique. Testing is one technique for gathering information which you can apply to a variety of decisions you wish to make, including student placement, diagnosis, mastery, and program refinement. Step 8, the Design for Data Gathering, is where the design of the actual test forms is accomplished and the strategy for administering the tests is

completed. The basic point of this paper is that when you are developing evaluation components for your own classes, departments, or colleges, the key testing instruments should be domain-referenced and not norm-referenced.

Domain-referenced tests are designed to measure performance on a clearly defined objective or set of objectives so that test scores are directly interpretable in terms of the objectives (or, in other words, the performance standards). At this juncture, I should explain why the term "criterion-referenced" is being supplanted. The word "criterion" has other meanings in testing and experimental design. It also does not convey the meaning that test exercises which measure performance on a given objective should be representative of all the exercises which could measure the objective so that we can infer learning and nonlearning from the test results with some certainty. A domain-referenced test, rather than determining how an individual stands in relation to other individuals in terms of overall achievement in a particular global content field, i.e., what percentile his or her score falls at, provides information as to which specific skills and knowledge an individual knows and does not know.

The distinction may be better made through example. Let me select a situation away from the paper and pencil tests we are most used to, and look at an area requiring psychomotor performance. Let's imagine you all out there are the examinees. The 1976 Summer Olympics are going to be held next summer fairly near to us in Montreal.

Our objective is to have each of you become a proficient short distance runner—that is, a sprinter. First, we are going to have a DRT testing program. The examiner might describe the situation and task to you as follows:

"This is an Olympic Stadium. Before you is a 400-meter track. Your overall task is to master the 100-meter event which is running down one side of the track as fast as you can. Our goal is to discover your strengths and capitalize on them and to discover your weaknesses so that we can improve your performance.

You will start at the north end of the track. The following skills are important and will be demonstrated to you, and you will be asked to practice them:

1. Position at the starting line
2. Instant of starting
3. The start movement
4. Leg movement while running
5. Arm movement while running
6. Body position while running

7. Means of acceleration
8. Staying in your running lane
9. Breathing
10. Endurance

We will work with you in groups gradually going through all ten skills and then putting all ten together into one sequence. There will be ten trials for each of the ten skills and ten trials of the final sequence of skills. It has been ascertained that this domain of skills and techniques, besides giving you a chance to participate in the Olympics, includes the major skills and techniques necessary for typical convention goers who attend meetings timed two minutes apart in widely separated hotels!"

After each trial the examiner will summarize your progress or lack thereof. For example, after a particular trial for all ten skills in sequence you might receive the following report:

"O.K., given that you have completed the trials on each of the ten separate skills and have completed two complete trials on the sequence of the 10 skills, I calculate you will make about forty percent of your widely separated convention paper readings on time. You are only fair on your fast starts. You need to work on coming out of the crouch position a little more quickly as it will cut down on the number of bruised ribs you will get from the knees of other dashing conventioners. You do very well on your leg, arm and body positions as you run. More practice should increase your endurance so that upon reaching your meeting destination you will not interrupt the already nervous paper reader with loud gasps for breath. Also keeping in your lane at crosswalks will undoubtedly add to the total number of paper readings you will be able to attend. Given the small sample of trials, I can't give you much more diagnosis. However, further trials should improve and extend diagnosis."

A situation for instruction in running as I have just described, provides a clear definition of the domain of tasks and gives you, the learner, much information for further improvement and practice. You can practice each skill separately and in sequence. Growth can be measured across time for each component skill and for the overall skill package. You are not competing with the performance of others. You have your goal specified: development of a series of skills for either entrance to the Olympic trials as a sprinter or as a person who attends paper presentations spaced closely in time, but distant in location.

We could set up other "Olympic stadiums" for conventional (if not convention-oriented) areas of learning such as English, chemistry, woodworking, physics, art. Perhaps you do so weekly.

Now let's look at the Olympic stadium with an NRT program.

Your examiner might say:

"Hi, conventioners! We are gathered here to compare your performance with that of your distinguished colleagues. Get down to the north end of the track and line up at the starting line. When you hear the gun go off, run as if your life depended on it down to this tape. The test will consist of fifteen trial dashes with ample time for rest in between. Each dash will be a different length as actual dashes would be for different convention meeting rooms."

The report of your score at the end of the test could be as follows:

"You scored at the 58th percentile with respect to the norm group of 500 community college conventioners who have attended conventions with at least two widely separated hotels where paper readings were held during the past five years. Thus, 42 percent of the conventioners did better than you in either getting to a meeting on time, getting there uninjured, or getting there without a general commotion caused by loudly gasping for breath. When compared to the norm group of all college faculty convention goers across the country, your score was at the 72nd percentile (probably because community college faculty are generally younger and more agile than four-year college faculty)."

Well, I think I have about run my example into the ground. So let me summarize. The NRT example does not clearly define the behavior desired. It is a global test of an attribute and results are given in comparison of one person's performance with another's. The DRT clearly defined the task and provided feedback in terms of learning and nonlearning of specific tasks or objectives.

Once again, let's focus on the definition of the two kinds of tests.

NRT—a test which differentiates among individuals on a particular attribute.

DRT—a test which measures an individual's performance on a specified task or series of tasks so that the test results are interpretable as to the learning or nonlearning of the tasks.

Listing the differences between NRT and DRT we have the following:

1. *Use of Tests*

NRT is used for placement and for prediction of success. For example, we could use a norm-referenced test for the selection of a number of individuals using a cut-off score for such programs as: (1) selecting students most likely to succeed in a block of vocational education training courses, (2) entrance to a highly competitive college, and (3) entrance to a job training program.

DRT is used for assessing student learning of particular, clearly defined skills and knowledges. One can collect information on monthly, weekly, or daily progress of an individual or a group on a particular set of instructional objectives. Tests results may be used to study the effects of instruction on individuals or groups in order to refine the program and diagnose needs.

2. *Construction of Test Exercises*

NRT test exercises must discriminate between examinees. Thus exercises are constructed with statistical characteristics which render them useful for spreading persons by their test performance on a continuum. For instance, an exercise may be desirable because it is generally answered by eighty percent of high scoring examinees and only twenty percent of low scoring examinees. We may begin with an objective for which items are written, but the value of the exercise is not judged on how well it measures the objective, only on how well it discriminates between high scoring and low scoring individuals. We tend in our own classroom tests to make "maverick" norm-referenced tests. We include trick exercises to separate the "men from the boys" or the "women from the girls," if you will. This is not wrong. But we should be aware when we are doing it.

DRT test exercises must be direct and representative measures of a given specified objective. Thus, the pool of exercises for a given objective must measure the objective completely and any sampling of exercises from the pool for test construction should be representative of the entire pool so that one can infer learning or nonlearning based on the test results. Item discrimination statistics are irrelevant.

At this point I should say that I am using the terms "exercise" and "item" synonymously. Any test question or test task of any type, I am terming an exercise or item.

3. *Relation to Content Domain*

Building on what was just said, this may be expressed as the question: How representative is the test of a content domain, or, of all the possible test items associated with the objective?

NRT relates to a generalized content domain such as reading ability, general accomplishment in biology, or even an hypothesized construct such as intelligence. It is not

specific in what it measures in terms of component skills and objectives.

DRT is based on objectives expressed in terms of specific, observable behaviors. The test is designed so that a student's score on the specific items written for the test are representative of all the possible items of the domain. If I wish to be sure that a student can add all combinations of two digit positive integers, I need to provide an adequate example of items which will represent that domain.

What this entire discourse is aimed at saying is that tests are vehicles for obtaining information for evaluation purposes, and by evaluation purposes I mean the assessment of student performance for use in improving instruction, redirecting instruction, and in judging student success in your course. Since NRT is concerned with placement, selection, and prediction, it has little, if any, use for evaluation design purposes. However, DRT is just exactly what systematic evaluation designs call for. It gathers data on specific objectives so that instructional decisions can be made by the decision makers, be they teachers, students, counselors, or administrators.

Let's take an example. Suppose there is an NRT test for veterinary science general achievement. Jerry Jones obtains a raw score of forty-six items correct which is at the 75th percentile level of performance. This means that Jerry scored equal to or better than 75 percent of the group used to establish the norm. If the group were composed of 20,000 second-year veterinary students, his performance ranks his proficiency in veterinary science in the upper 25 percent of the students *as measured by that single test*. This does not mean that Jerry is or should be performing at a certain level in class. It is just one indication (provided the instructor accepts the test as a good measure of veterinary science as he teaches it and that Jerry is in fact a sophomore in veterinary science) that Jerry has a better than average knowledge of veterinary science (as defined by the test).

Of what immediate instructional value is this test result to Jerry's instructor? If Jerry is doing poorly in veterinary science, the instructor should carefully look for reasons other than non-capability and nonreadiness. If Jerry is doing well, the test confirms this.

NRT is of limited value for evaluation designs. It does not relate very much to the classroom. Its items are not directly related to instructional objectives and, thus, its results are not useful for making decisions about student success in a course, suitability of program components, or emphasis of instruction. In short, NRT is

not designed for making daily, weekly, and monthly decisions for the management of learning. DRT is. It relates directly to needs for information for program effectiveness, student diagnosis, and student progress.

The questions may well be asked: What is a good domain-referenced test? What should it contain? The remainder of this presentation will be devoted to describing various designs of domain-referenced tests and the decisions for which they provide useful data.

A short overview of the kinds of information needed by educational decision makers is provided as an introduction to ways of using and designing domain-referenced tests.

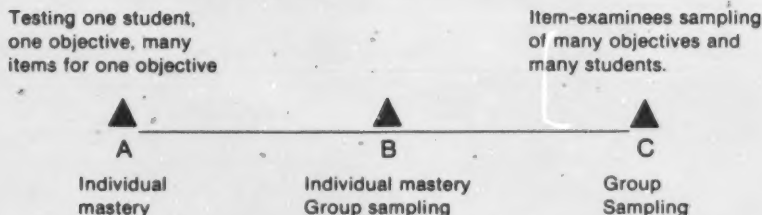
Using the approach described by Dr. Gorth, our first step is to identify the enterprise, the focus of the evaluation. Let's select the student as our enterprise. Assuming that Steps 2-5 are completed (Resources, Decision maker, Goals, and Objectives) we come to Step 6, the identification of the decisions to be made about each objective. I'll set myself up as the key decision maker.

Typical decisions I wish to make may be one or more of the following: (Remember that these decisions are focused on the evaluation of the individual student.)

1. *Placement within the curriculum sequence.* The key word here is "within"; thus, a DRT approach is called for. NRT may have been used to select students for course entrance but DRT is the tool for accurate placement within the course.
2. *Determining preknowledge* which might be stated as the question, "What background or readiness does the student have for taking my course or a particular segment of my course?"
3. *Determining immediate post-instruction knowledge.* This is what we traditionally test for. We teach; we build a test hopefully measuring what we hopefully were teaching, and we test after the instructional sequence is completed.
4. *Determining retention.* Weeks after instruction is completed, does the student retain the learning?
5. *Determining overall achievement.* Given my course, what success has the student finally achieved?

What DRT test designs are necessary to produce information which will be useful for these decisions? All five can be handled with a similar testing design. To show you this, let's conceive of all DRT designs as falling on a continuum. One pole of the continuum focuses heavily on a single instructional objective and a single individual, and the other on a wide range of objectives perhaps covering an entire course or even a year's sequence and a group of individuals.

FIGURE 1
Continuum of Test Composition.



Since all five decisions relate to individual student mastery of a particular domain of content we could have, in the simplest case, a test consisting of several exercises, perhaps as many as 5, 10 or 20 which measure the performance of a single objective. Depending on the number of objectives we would like to test, we could limit the number of test exercises per objective so that a single test form could measure several objectives. Thus, within-course placement, preknowledge, immediate post-instruction, and retention could all be accomplished with similar formatted test forms which are testing mastery of several objectives. Overall achievement may require several test forms. We might also combine objectives which are relevant to several of the five different decision categories into one test form to achieve an economy of testing. The focus of mastery testing is on obtaining a reliable indication of student knowledge. This requires using several exercises all related to a single objective so that the student has a chance to react to an adequate, representative pool of exercises for a given objective.

For example, suppose you are testing an individual's performance on five objectives. You may have a twenty-five item test containing five pools of five items each related to a single objective. To ascertain achievement you may set a performance standard (i.e., mastery standard) of four out of five items. Thus, a student must correctly respond to four out of five items for each objective in order for him or her to have shown mastery of the objective.

Now suppose that the enterprise is not a single student but a group of students—such as the 500 students in introductory English Composition. Again, using myself as the decision maker, I identify the following questions of interest to me.

1. *Are the objectives of this course selected for the target population appropriate?* That is, will I be teaching objectives which will be new to the students and which will transfer from past learning?

2. *Are instructional materials teaching students of different abilities equally well?* In other words, is there a link between how well students are learning and the instructional materials being used?
3. *Are the instructional objectives sequenced appropriately?* Is the sequence of learning logical and sensible?
4. *Is there sufficient redundancy in the curriculum to maintain learning?* Do students retain learning of course components across the course?

When you wish to gather information on a group of students, you are probably different from other instructors. Most of us give tests which focus on individuals, and rarely do we systematically gather information through testing which will help us improve our overall program. I am speaking of formative program evaluation, which focuses on the suitability of what you are teaching, when you are teaching, and how you are teaching. You may not see a need for program evaluation, or you may not see how tests can give you information for program evaluation. Many people don't. They generally have arguments such as the following:

Argument 1: "I have a certain core of introductory material that I must get across. I have fifteen lessons to give and each one is laid out in unchangeable perfection. If students are not interested, they don't have to take my course."

There is a difference between non-interest and confusion. Maybe all of us, no matter what we teach, would do a better job if we received and used test results related to the four decision questions I have raised.

Argument 2: "I don't have time to test. I see the students only twice a week."

Good tests can be very short and easily corrected. In fact, most DRT tests should be relatively short for the best testing situation to have students do their best.

Argument 3: "I don't get to know my students well enough to make decisions about alternative instructional approaches."

Maybe you don't, but you can test to discover when you are being overly confusing or making a large jump from one point to the next. Even if you cannot go back this year or spend more than a few

minutes going back, next year you could be aware of the problem and adjust instruction accordingly.

Looking back at the test composition continuum (Figure 1), Pole C is the extreme testing format for obtaining group information. It is a form of item-examinee sampling, and there may be several test forms, perhaps not parallel in the norm-referenced sense, but randomly parallel based on the objectives being sampled on each form. Let me illustrate. Suppose you teach "Basic Chemistry for Nursing." You identify thirty key objectives which are the heart of your course. You develop a pool of test items—perhaps five to ten for each objective. You can then build a series of randomly parallel tests which will monitor group progress (and thus course progress) across the semester on these thirty items. You decide to randomly assign one test item from each item pool to each of five test forms—giving you a series of five tests each of which contains thirty items, one item for each objective. One test form per month can be administered to each student with all test forms being used at each administration, each student taking a different one of the five forms each time. (See Table 1.) Group results by objective can be computed and either graphed or be placed in a table of percent of items answered correctly by the group at each test administration for each objective. The trend of the percents over time could be checked against your expectations for the group. (See Table 2.) If objectives are being learned when they have not been taught, or not learned when you thought you had taught them, or forgotten quickly after you thought they had been thoroughly taught, then you can look back for the causes and can make a decision about a remedy.

Other designs are possible. The midpoint between Poles A and C, that is, Pole B, may be thought of as a test with both mastery and group data components. Suppose we construct a test of thirty items. Twenty of the items measure the four current objectives covered during the past two weeks, five items per objective. The other ten have other functions. Five could sample group performance on objectives being retention tested from previous course segments and five could be pretesting the next course segment. The test content could slide for each administration across time so that the objectives represented by the three testing phases of the test—pre-test, immediate posttest and retention test—would change phase from one test to another. This design would provide information useful for making all nine of the instructional decisions mentioned.

In conclusion, I would like to remind you that testing does not exist separate from an evaluation design. The usefulness of the design is up to you. You may cringe at the effort involved in producing or selecting learning objectives and test exercises, but good

exercises should be reused, and with an item-examinee sampling design, they may be reused many times during a single semester.

Going back to one of the comments I quoted earlier, "You should not teach for the test," we find that, in most instances, we do want to teach for the test with DRT. I don't mean that we should teach the items or exercises on the test, but that we should be teaching the objectives. When a student incorrectly responds to a test exercise, it is not the exercise we should focus on with him, but the objective to see where his problem in learning lies.

TABLE 1
Students Grouped for Item-Examinee
Sampling Testing

Student Subgroups	Test Occasion				
	1	2	3	4	5
A	1*	2	3	4	5
B	2	3	4	5	1
C	3	4	5	1	2
D	4	5	1	2	3
E	5	1	2	3	4

*These numbers designate the different test forms.

TABLE 2
Trend Data Generated from Item-Examinee
Sampling for One Instructor's Students

Instructional Objective	Monthly Test Occasion				
	1	2	3	4	5
101	.10▲	.50	.80	.90	.90
102	.50▲	.75	.85	.90	.90
103	.10	.10▲	.75	.95	.95
201	.05	.10▲	.85	.85	.85
202	.05	.50	.90▲	.90	.90

*Proportion of students answering items for the objective correctly.

▲Point that instruction occurred.

Fair Measurement And Higher Education

EVA L. BAKER

At long last, concern has been raised toward improving the practice of instruction and evaluation in higher education. Community colleges, four-year institutions, and multicampus universities are all, in some manner, confronted with the same sorts of instructional issues that have plagued elementary and secondary education for two decades. Given the aroused attention in these areas, educators at the college level must decide if they are willing to undertake productive action. Across all aspects of education, the topics of evaluation and measurement are commanding a growing respect, and it is to these related areas that this paper is addressed.

While students have alternately muttered and ranted about the unfairness of testing for years, their major concerns at the higher education level have been directed to perceptual rather than technical matters. Because the student groups in higher education in general, and community colleges in specific, are comprised of older people, reactions to the process of measurement are usually mixed. Objections are generated partly from the fact of adulthood and the sure realization that artificial measurements are rarely a regular expectation in the world of work. When school has been interrupted by job experience, the sense of dehumanization can grow. Students also object to tests because of the usual relationship of tests to the practice of assigning grades for work accomplished and the consequences of grading may be awesome with regard to self-esteem. Fear of failure as well as need to excel engender the highly competitive or at least anxiety-ridden climates that surround many institutions of post-secondary learning. The reality of the effects of grades in the pursuit of further education or professional placement increase the pressure to compete and student dissatisfaction with the entire enterprise. When students have, in the past, voiced their concerns about testing, they have usually not addressed the problems of improving measurement but in getting rid of the whole annoying process.

Educators have, on the other hand, treated the measurement requirement with more acceptance (perhaps because they ran the gauntlet themselves and certainly because they are not presently required to be tested). The public has had evaluation rhetoric repeatedly passed before their eyes, and legislators at state and national levels are applying more pressure to educational institutions

to determine the effects of their programs on students. It seems that the evaluation and measurement questions will not disappear. In this paper, I would like to discuss what I see as the future in the area of test design and implementation, the most promising ideas in instrument development, and then some specific procedures that could be used to implement a program of measurement improvement either for a given department or on an institutional basis. Certainly there will be no guarantees, but the focus of this presentation will be strategies that we can use to incorporate better forms of evaluation and measurement in community colleges.

Contexts

In trying to improve measurement in the college setting, we immediately encounter obstacles of enormous proportion. The first is the fact that faculty have been insulated from serious criticism about their test development. Protected, because we have made the common assumption that subject matter competency equates with measurement competency, the faculty member may not recognize deficiencies in the level of present practice in measurement. College faculty, by and large, accept the premise that one who is recognized as an expert in the field will be able in some inevitable way to evaluate others' abilities. Reliance upon such a superstitious view is not unexpected when any of us recalls our own college and graduate school experience. Degree award often seemed to be on the basis of chance and chosen on good guesses and hopes rather than because we could demonstrate specific abilities, or even, for that matter, knew that we had them. Moreover, the error of assuming that subject matter competence implies teaching competence has been observed in institutions around the country. While at the multicampus university it may still be the case that high scholarship in research emits an aura of teaching competence, the falseness of such a view can be substantiated by oblique evidence. John Goodlad found, in reviewing the large scale curriculum development projects, that although training of teachers was devoted toward making them more competent and comfortable in the subject matter, the failure of the new public school curricula could be attributed in large measure to a failure of pedagogy. The teachers did not know how to teach the new concepts that they had learned. Since instruction is a more open and observable arena than measurement, one can observe teaching, inspect required materials, and discuss instructional events with students and colleagues. *En masse* boredom or anger suggests that complaints about poor instruction may be justified. Even though such judgments focus on the process of teaching (what the teacher is doing) rather than the outcomes of

teaching (what effect the process has had on participants), instruction is at least a series of events available for review.

On the other hand, the test or measurement instrument is more difficult to judge. Good tests may not look much different from poor tests, assuming that gross format errors have been avoided. All tests have questions or problems, they are usually printed, in black and white or at worst, ugly purple, and they may have the added attraction of answer sheets or the requirement of blue books. But the visible manifestation of a good test is not the way the test looks, but usually the way the test has been developed. Before the specific issue of test design can be attacked, the various reasons for testing should be explored. There seems to be unparalleled confusion about the reasons for measurement in educational settings: Among possible purposes, let me review a few:

Measurement is:

- to have a basis for assigning grades (meeting an institutional requirement)
- to provide personal feedback to the instructor about the success of a course
- to enable us to understand what our students know so we can communicate with others
- to provide a basis for allocation of funds for program initiation, improvement, or reduction.

It is clear that some of these purposes are overlapping, but some are quite distinct and suggest very different ways of approaching the testing problem. For instance, let us suppose that the major purpose of measurement is to provide grades to meet an administrative requirement and that we are operating under the illusion (or requirement) that we must have a wide range of grades, not all "A's" but something approaching a normal distribution. Our measurement procedures could be relatively simple. We need to develop a procedure that spreads out the student group and permits us to discriminate among them. So a difficult test related to only a small part of the course might serve our purpose. Furthermore if the test is *truly* only to meet administrative requirements, we don't need to worry about reliability across test administrations. That is, if a person would do well during one testing and poorly during another, we are not concerned, since our main objective is to develop a range of per-

formance, so we can give grades. Naturally, this is an extreme view, but very probably represents much of what goes on in the name of measurement in postsecondary education.

If on the other hand, we were not so concerned with grading, and merely wished to describe what our student group could do, we could design a comprehensive test that sampled all important aspects of the course content and skills. We would want that test to be reliable for a given student set and might need to modify the test as we change the course. If we want to use the test to determine if a new program was an improvement, we would need to construct the test so that it would be reliable across time, so that we could compare year two with year one. If we wanted our tests to help us specifically to improve our instruction, that is, for the purpose of formative evaluation, then test items would need to be diagnostic. We would want the items written keyed to specific portions of the course so that we might improve particular activities or assignments. While there are certainly many more reasons for which test development would be modified by purpose in an instructional context, I think the overriding concern in measurement design is that the test be regarded as a fair measure of the course by both faculty and students.

Faculty must accept the measurement as a fair indicator of ability in the domain of concern. No faculty willingly sacrifices significant attributes of content to technical efficiency. We should not ask students to underline complete sentences in paragraphs when we really want to know if students can generate their own sentences.

As important, the measurement procedure should be understandable to the students. For too long, measurement has been purposely shrouded in mystery so that students have sometimes accepted the "greater wisdom" of the test designer even when logically the test seemed, at best, peculiar. I am not one who believes that public education has to compete with multimillion dollar media programming to win student acceptance and approval. But as going to college becomes more an act of will rather than an expectation, and satisfaction of students occupies growing concern in a time of competition among institutions, I think that students must have the sense of fairness in measurement reinforced. People can accept poor grades, and routine or even witless instruction, if they feel during testing that they are being treated fairly.

Newer procedures in the improvement of test design are important because they can realistically support the idea of the fairness of a test.

Without getting too detailed or technical, let me describe to you

one of the improvements in test development that has been evolving in the last few years.

Without doubt you have heard of the ideas of objectives-based tests or sometimes called criterion-referenced tests. These are tests which, in the simplest terms, measure the outcomes that the instructor plans. These tests are becoming increasingly important when the purpose of measurement is evaluating programs rather than simply discriminating among students. Most commercially prepared achievement tests to date are not of the objectives-referenced type; yet large test development agencies such as Educational Testing Service are now working in the criterion-referenced test area.

The logic of criterion-referenced testing is simple: your test measures the desired outcome. It would seem that almost anyone could develop a test that matched clearly stated goals or outcomes. Unfortunately, when reviewing tests developed by instructors we have found an enormous variation in the kinds of test items one instructor thinks is a valid representation of a particular objective. This disparity in interpretation becomes important when students are expected to be certified competent in a particular area, such as composition. We would not wish students to have an especially difficult or easy time of it as a consequence of measurement vagaries. Such a situation would not contribute to a perception of fairness, especially when students wish to understand clearly what is expected of them. In addition, measures that varied in difficulty from year to year with the same instructor would not be easily justified. The main contribution of objectives-referenced tests, so far, is the guide they provide for determining whether the actual test items used in the measurement is a veritable measure of the goal. For instance, if students are asked to identify the theme of a one-act play in an objective, and in the test are asked to write an essay identifying theme and justifying the author's choice of tone, then the measurement goes beyond what is called for in the objective and is not fair.

A simpler example might involve objectives that call for students to analyze the great movements in American history, yet test students using a multiple choice test of information.

A further refinement of the idea of criterion-referenced tests was developed about eight years ago, largely under the influence of Wells Hively. Hively terms his refinement, "domain-referenced achievement testing."¹ (The term domain should not be confused with the use of the word in connection with the cognitive and affective taxonomies of education. Rather, we are using domain in a common language way.) Domain-referenced achievement tests try to explicate the limits of any objective, so that any number of test

items could be considered fair or representative measures. They are based on a clear statement of the test domain. This domain is used to guide one or more test designers to develop the measures. It serves as an operational definition of fairness.

The elements of domain-referenced achievement tests include:

- A. The statement of an objective or general response description
- B. Limits of content
- C. Criteria for adequate response
- D. Format constraints

I'll briefly review the requirements of each of these points just to identify the main improvements that domain-referenced achievement tests afford the test writer.

Statement of the Objective

The statement of the objective is stated in much the same way as any model of an instructional objective. The outcome is stated in terms of learner responses that might be observed. (This element is the same as used in criterion-referenced test development.)

Content Limits

Stipulating content limits requires that the instructor describe the range of content that can be fairly sampled in the test item. The effect of content limits are two: first, an objective is not likely to be stated in terms of a single item. For example, we won't teach students to take photographs with a Nikkormat camera; we will teach the students to use any single-reflex lens automatic camera. Secondly, we won't be able to get by with vague allusions to the content areas that we will cover. We cannot say that the student will answer questions about various propaganda techniques in mass media presentations. We will have to describe the kinds of questions they may be asked and the range of media presentations that might be queried. We can't say that students will solve word problems in which common computations are required; rather we will define the limits of complexity or simplicity of such problems in terms of the numbers of operations and the complexity of the prose problems.

The use of content limits represents a significant departure from present levels of objectives-based or criterion-referenced test development. And, as you can see, they have important implications for instruction. When we describe a legitimate testing domain as

single-reflex lens camera, rather than a specific manufacturer, we are required instructionally to take account of a range of cameras, for example, the variety of manipulation to control shutter speed and film rating on different cameras.

Criteria

The term criterion has long been used in educational terms to mean standards of adequacy. In much postsecondary education, such standards are rarely articulated. An essay is good or it isn't; a term paper is an "A" or it is not. The criterion section of a domain statement requires that one specify the bases against which responses will be judged successful or not. The problem is taxing but becomes more complicated when we think of multiple choice tests. The implications of the criterion section of a domain statement is that the range of fair wrong answers will also have to be given. For example, in a test in which a student is to select the reason an author used a particular incident in a short story, the kind of distractors or wrong answers given could greatly influence the level of instruction one gave. If wrong answers were provided as distractors which included at least one correct element, then we would be working for a higher degree of learning than if the wrong answers were all far off base. While it is difficult to provide enough examples to make sense to people with diverse subject matter orientations, the explication of criteria of adequacy is the second major improvement that domain-referenced achievement tests offer.

Format Constraints

The fourth element of a domain statement is the test format. In this section the test designer stipulates the form the measurement is to take. If paper and pencil tests are to be used, the format would describe what additional information the student could have available, how many distractors would be presented, and what the directions would say. If the test were not limited to paper and pencil, the format would provide the directions that the test administrator would give, or describe the laboratory setting or the performance context for the measurement.

Domain-referenced achievement testing is probably one of the most discussed innovations in the area of test development. Further description of the process of domain specification can be obtained from *Domain Referenced Testing* (Hively, 1974).

Practically, the adoption of such a procedure would enable a faculty member to:

- have test items written by colleagues as well as himself so to provide a less idiosyncratic interpretation of the field.

- permit the generation of a range of acceptable measures, so that test development need not be rushed at times of impending examinations nor be vulnerable to the horror of stolen examinations.

- provide a basis for communication of program outcomes for students and colleagues.

- encourage tests to focus on generalizable skills rather than idiosyncratic ones.

- provide specific information useful for instruction improvement.

(Deficiencies could be attributed to inadequacy of preparation on, for instance, the range of content of examples provided, the specific instruction regarding criteria, or heightening the discrimination among confusable but wrong answers.) When test domains are well explicated, there also may be a small, faint hope that, for certain courses, instructional prowess will be judged by how many students do well, rather than by how many students do poorly.

Domain-referenced achievement test procedures are difficult to employ. In projects at UCLA, we have had to exert more energy into the creation of domains than we ever imagined. Yet the benefits at this point seem to be worth the effort.

Better Evaluation Data

In addition to improving the basis of construction of measures of the actual achievement of students against the goals that we have stated for them, colleges could make better use of other sources of information for the evaluation and modification of their program offerings. In this regard, many institutions have developed student evaluation systems for providing information on student reaction to instruction and instructors. These forms are completed at the close of every term. The data are usually fed back to appropriate departments and are sometimes published, in their most brutal form, in handbooks distributed to students at the time of enrollment. At many institutions, student ratings are used at the time of faculty promotion and advancement to estimate the instructor's teaching ability.

The use of student reaction measures is an important aspect of program evaluation that should not be overlooked. Unfortunately, the design of most of these instruments is very unsystematic. In an attempt to be comprehensive, many of the questionnaires require

ratings to 50 or 75 separate questions. When one multiplies the number of courses a student would ordinarily take per year times these questions, it is easy to predict that in a short time attention would wane and fatigue would dominate the reactions obtained.

Certain simple procedures might be developed to help make student reactions a more informative source for program improvement. First, questions might be segregated into non-overlapping categories. Students might be asked to react to the importance of the goals and to estimate their personal learning from the course. These sorts of questions seem to me to be of greater importance than assessing whether students liked a particular instructor's lecture or discussion style. A second section might ask for recommendations regarding areas for improvement. In my courses, my students are extremely explicit in suggesting activities that might have helped them acquire the course goals and in identifying those procedures that they thought inhibited them. A third section might attempt to assess outcomes (other than those anticipated) from the course. Students might be asked to indicate the best and worst part of the course. In addition, they might be asked to explain any application of the content that they found outside of the particular course being evaluated.

Reducing the broad student evaluation base to a very few questions that have clear implications for improvement should increase the efficiency of the process as well as reduce student fatigue. In addition, sampling procedures might be employed so that only a proportion of the students respond to any given course evaluation sheet. It might be more meaningful, for instance, to ask half of the students to respond to a questionnaire immediately following a course, and the other half six months later. Such a delayed administration would in some small way help us deal with the nagging concern that our testing periods always come too close to the end of instruction. We usually collect such attitude information just at the wrong time, when anxiety from examination rigors is greatest and perspective is weakest.

We could also employ evaluation tactics that are somewhat less conventional. All people are wary about questionnaire information, particularly when it is referenced to a single individual, such as a faculty member. The administration of the campus might sample among students and ask them to comment, across departments, about their best and worst campus experiences. They might also ask for recommendations for improvement. This information might be gathered in questionnaire form, or perhaps students might be asked to prepare two minutes of comments and record them into a cassette audio tape recorder.

Faculty too should be surveyed to assess their sense of satisfaction in teaching the courses assigned to them. Somehow we often ignore the primary participants in any instructional effort, the teachers, and miss useful information. It is usual to expect, for instance, that faculty will prefer to have small rather than large classes. In my department, however, a few of the faculty expressed dissatisfaction with small classes and their preferences for large. If such information had not surfaced, program reorganization would have proceeded along very different lines. When we determined that large class enrollment did not stupefy all of us, then we were permitted to vary our enrollment limits more, so that some students would be working in groups of twenty-five while others were in groups closer to one hundred.

In some way, we wish to involve both students and instructor in expressing their sense of satisfaction with the learning-teaching roles in which they are cast.

So far, I have outlined some concerns in the area of measurement, identified domain-referenced testing as a potentially useful procedure for improving achievement testing and described some problems and solutions in accumulating data about student and faculty satisfaction. If you're interested in self-evaluation, the reverberation of such remarks at this point should be, "these are wonderful ideas! How can we do them?" For the final section of this presentation, I would like to identify some procedures that you might use to attempt to implement changes in the basis for the evaluation of instruction.

To Improve Evaluation—Really

Each of us should be instantly disabused of the notion that it will be easy to change our evaluation practices. Particularly in the area of the measurement of achievement, we are running smack into a number of serious obstacles. First, there is the issue of personal responsibility. Sometimes masquerading under the title of academic freedom, instructors will exhibit reluctance to make public the basis upon which they evaluate their students. For them, evaluation processes may be too private and too vulnerable to be displayed. Second, we have the realities of academic life. We wish to accomplish the change within the steady stream of other responsibilities, such as advising students, teaching, devising new courses, preparing again for old ones, meeting administrative and committee responsibilities, and always hoping for a little time to think. Next, we must recognize that such effort to improve evaluation is not free; it is not something that can be casually added onto the existing load at no institutional cost. Finally, we must face the issue of incen-

tives. The consequences of changing the basis of measurement will not be immediately observable. It is not the same as developing a film that can be shown, or a textbook that can be read and hopefully sold, a Batik that can be praised, or even an attractive brochure advertising a program. The outcome of evaluation improvement is likely to be manifest over a longer period and in no single dramatic way. Thus, not only do individuals need incentives for participating in the venture, so do institutions.

Specifics: A Process of Development

In order to have a chance of success, it appears that improvement in evaluation must be treated like improvement in any other sector of education: as a process of development rather than a fiat. Among the basic tenets of empirical development in an institutional setting, paramount is that one must keep control of the process. Some specific procedures that you might use are as follows:

1. Do not attempt to implement new evaluation procedures across the board at your institution. Involve relatively few units in your campus.

2. Do not expect to use these procedures with all kinds of courses, or even for all courses taught by a single instructor. Instead, select a small number of courses where the potential number of students affected is high and thus the impact of even a modest improvement is great.

3. Attend to the issue of institutional incentives. The institution will not support financially or maintain interest in measurement changes where it encounters great resistance from powerful faculty. Attempt to involve prestigious faculty or usually recalcitrant departments very early in the design and participation of this activity. Success with these individuals will increase the credibility of the effort, and make it easier for a larger section of the campus to be encouraged to participate. In other words, use elitist drives where you find them!

4. Attend to the issue of incentives for individual involvement. Can a campus-wide handbook be developed detailing the procedures followed, for which the participating faculty might be coauthors? Can the process be disseminated to similar departments at other comparable educational institutions?

5. Provide a means for participation within the press of other requirements. Rearrange course assignments so that involved faculty receive institutionally supported released time to engage in the process. If neither course release nor summer salary is possible, consider reducing the instructors' clerical responsibilities by

arranging for help. Often local high schools are willing to provide work-study students to assist in clerical assignments at no cost to the employer.

6. After development of new measures, secure appropriate collegial review and approbation. Permitting involvement at every step of the way can help you to avoid eleventh hour objections.

7. Beware of false technology! Keep computer analyses out of the hands of the faculty at first. After learning more about the process of domain-referenced testing, one will see that the usual test analysis programs available to analyze student performance data are not likely to be useful. If some well-meaning person attempts to provide inappropriate statistical information, faculty members may become intimidated. Faculty will need to be desensitized of their fear of numbers, if they have any. But elaborate routines generated from magical computers may inject the entire activity with a sense of unreality and lack of control.

8. Don't expect to complete the task and implement new measures once and for all. In the development of evaluation measures, we should not expect a revision every quarter or year, but a regular interval for review perhaps on a three-year basis should be anticipated.

9. Share what is produced across the faculty. Wait until some level of satisfaction is felt by faculty participants, then give them the mission of helping to train other faculty in the task.

10. If you cannot envision implementation on a department-wide or institutional basis, find a friend (you'll need at least one) and start the process yourself.

Conclusion

While no one sensible would minimize the difficulties in internally generating and maintaining pervasive change in educational institutions, nonetheless, the area of measurement is a seductive one on which to focus. Good student evaluation directly affects the way students perceive the institution and their role in it. In addition, careful development and use of measures can influence positively and explicitly the kinds of instructional programs we mount. The cost of such efforts is comparatively low when contrasted with capital costs associated with purchasing new hardware. The area of measurement has an expanding technology that can be made available to a wide segment of the educational community. With a relatively small investment of resources, community colleges could lead the way for a revolution in the applications of technology to further individual satisfaction.

Methods Of Evaluation In Non-Traditional Settings

RICHARD G. ALLAN

Vice-President for

Instructional Development

National Evaluation Systems, Inc.

Introduction

Community colleges represent some of the most innovative and pace-setting types of institutions in American education today. They are a viable alternative to four-year colleges, immediately going to work or entering the service. Two-year colleges, as a real alternative, began to make a major impact on the general public consciousness during the late 1950's and early 1960's when the California system received widespread publicity. Since that time, community colleges have literally exploded, both in terms of quantity (institutions and students) and program offerings.

In the past, the now distant past, post-secondary education was generally considered synonymous with four-year colleges or, in some cases, some sort of technical training. This narrow concept of post-secondary education is obviously outmoded. Community colleges offer a wide range of programs. The more traditional types of programs, such as liberal arts, leading to a four-year college, and technical training or career education programs, such as nursing, secretarial, or laboratory technician programs still exist. However, a whole host of new and creative programs are being developed. These new, "non-traditional" programs are mostly in response to efforts by the colleges to meet community needs, whatever these needs may be. The past decade has witnessed a great deal of thinking, writing, speechmaking and consequent implementation of needs analysis and needs assessment techniques to meet community needs. Consequently, new student populations, new types of program goals and new techniques to accomplish program goals have resulted.

Now that these new, non-traditional type programs have been instituted, a question must be posed—how does one evaluate non-traditional programs?

In order to deal with the question of evaluating non-traditional programs, it is necessary to first define the variable involved. The next section of this paper will define how I use various terms related to the above evaluation question.

There are a series of meanings and/or definitions of evaluation. The one I tend to favor is "Evaluation is to provide data for decision making." This definition has proved useful and workable in my experience. In essence, the evaluation process includes identifying who the decision maker(s) is/are, finding out what kind of decisions they intend making about the program of interest, and collecting the information that will lead to informed decision making. This process has been outlined in more detail by Dr. Gorth in his earlier presentation. I believe that this process will prove useful in evaluating the non-traditional programs that are being developed at an ever greater rate by today's community colleges.

Hypothetical Programs

As a first step in considering the evaluation of non-traditional programs, I will suggest a distinction between traditional and non-traditional. This distinction is offered solely for utilitarian purposes related to this presentation. The only purpose of making the distinction is to draw some boundary lines for purposes of identifying what programs will be considered and what programs will not be considered when looking at evaluation methods. A traditional program is one that is conducted in a classroom with a teacher as the basic dispenser of information to people generally considered "college students." This would include all lecture classrooms, large or small, and many seminar classrooms where inductive type instruction is occurring. A non-traditional program is one that is conducted using types of instructional approaches other than a group-paced, classroom-based, teacher-oriented instructional organization. The non-traditional program is quite often in subjects that are entirely new, i.e., not in the curriculum ten years ago. However, non-traditional programs also exist for teaching traditional course content to what has long been the expected student population at a community college.

Rather than always operating at an abstract level, let's look at some hypothetical non-traditional programs. Then we can talk about how to evaluate these hypothetical programs.

Program 1: Dietetic Technology Program

Diane Dietician has recently developed a Dietetic Technology Program at Holy Oaks Bluff Community College. This is a career-oriented program. It is not in a new content area, yet the course does contain elements of organization, development, and instructional strategies that fall under my previous definition of a non-traditional program.

The population in the program consists of women aged 18 to 54. They are all high school graduates. Some have a bachelor's degree and are enrolled in the program because they feel they will have a higher job opportunity potential. Men are not excluded from the program; they simply have not chosen to enter.

A needs analysis was conducted with the advice of members of the local Educational Research Center in a nearby state university. Program objectives were developed based on the needs analysis.

Program elements include courses taught in other departments, such as psychology and chemistry, courses taught in the Dietetic Technician Department and field experiences.

Instructional techniques include learning activity packages (LAPS), lectures, seminars, reports and field activities under the supervision of personnel in these field institutions.

Program 2: Minority Construction Contractor Program

Joe Steamshovel, at Man-o-War Community College, organized a course for construction contractors. This is a career-oriented program for a new population. Most of the courses are of a traditional nature.

The population in this program are all members of a minority group in a medium-sized metropolitan area where there has been a large influx of minority groups in the last ten years. The members of the program are all men, although women have not been excluded.

The program was developed as part of the efforts by the Community College to identify new and emerging needs of the citizens in the community. The program content was decided on, in large part, by the college.

The program consists of two major elements. The courses are, for the most part, traditional. They consist of accounting, management, estimating, etc. There is also a program element designed for self-support, confidence building, and awareness on the part of the students.

Program 3: Retiree Program

This program was developed by the Shady Glen Community Col-

lege for persons who had retired or were about to retire. The genesis of the program was an effort, recently instituted by the college, known as Community Outreach. Many aspects of the program are non-credit, interest-oriented activities.

The population consists of men and women who have either retired, or are soon to retire. The college is in a small metropolitan area, so program members seem to form a rather tight-knit group, able to interact well.

The program design is a collaborative effort by program directors at the college and members of the program. A council, consisting of members of the community and the college designed the program. The council now consists of members of the program and members of the college.

There are three basic elements to the program. The first element consists of courses and experiences for pre-retirees. The purpose is to ease the transition from a psychological and financial standpoint. The second element consists of activities and organization for activities to make use of retirees. These activities include instruction in crafts or child care, so that program members can return to centers for the elderly and work with their peers (or elders) or act as aides at child care centers. The third element consists of regular academic courses with components for the elderly. For instance, a sociology course made use of retired members of the class to take a senior population survey in remote areas of the state. The program members were able to make good use of the time on their hands, plus provide some human contact with elderly folks in remote areas who seldom came in contact with other people.

Methods of Evaluation

At least two more general areas must be considered prior to discussing the evaluation of the non-traditional programs described above. These areas include methods of evaluation and decisions to be made relative to the programs.

Rather than presenting an exhaustive list of methods of evaluation, I am proposing a series that are generally known and understood by most teachers and administrators. These evaluation methods are reasonable to use in the evaluation of most programs and, when used in combination, will provide sufficient data for informed decision making for most programs, even non-traditional programs.

1. Norm-referenced tests—These tests provide comparative data on students. The results allow for decisions to be made when one student is compared to a group of students or when a selection process is called for. For instance, if there is room for only a percentage of a certain population in a program, and the desire is for the most capable students to enter the program, then a norm-referenced test would be used.

2. Criterion-referenced tests—If you wish to measure a student relative to an absolute standard of some sort, then criterion-referenced tests should be used. A criterion-referenced test will tell you exactly what a student knows or can do. It relates student to content knowledge rather than comparing students.

3. Simulations and role playing—While there are differences between simulation and role playing, most classroom teachers would have a difficult time distinguishing between them as they are generally used in a classroom. Therefore, they are treated together for purposes of simplicity in this presentation. A simulation or role-playing exercise requires students to interact in some sort of contrived situation. Usually the situation presents a particular problem or issue to be played out by the students. For instance, a nursing student may be told to care for a patient exhibiting any one of a number of symptoms. Then the patient (usually another student pretending to be a patient) is told what symptoms to display. The nurse's performance can then be analyzed to determine if the proper procedure was followed for the symptoms displayed. Simulations are good to use when human interaction is an integral part of the process.

4. Questionnaires, interviews—Almost no one proceeds through life without filling out many questionnaires, checklists, or survey forms, and submitting to interviews. They are familiar enough to need no further explanation.

5. Unobtrusive measures—Unobtrusive measures are continually used in an informal manner by teachers and administrators, but are seldom formalized into an evaluation procedure by these same teachers and administrators. Usually, the unobtrusive measures are channelled into "gut-level" reactions or "vibrations" or some other semiconscious type of internalized feeling. Unobtrusive measures relate to collecting information on the subject (person) that he/she is not aware is being collected. The attentiveness of students in class, checking to see how many students have

read assigned readings in the library and noticing whether class or program enrollment is increasing or decreasing are unobtrusive measures that we have all applied at one time or another.

Decisions

As was mentioned above, the purpose of evaluation, as I am using the term, is "to provide data for decision making." So, prior to designing an evaluation for a non-traditional program, one must identify the decision maker and the decisions to be made.

The primary decision maker I will focus on is the teacher. Both administration officials and students are prime decision makers in any college program, but then decisions tend to be extensions of teacher's decisions on a smaller, more personal level, or on a larger, more global level.

Some of the decisions teachers must make are:

1. What should the content of the program include?
2. What is the best instructional strategy to use?
3. Were the objectives of the program learned?
4. Did individual students master the program objectives?
5. Were the students adequately prepared for "X" by the program? ("X" means whatever the program was designed to do.)

These five decisions represent a small subset of the decisions teachers must make continually. Yet, they are among the most important decisions made by teachers. Quite often these decisions are made based on less than satisfying data.

Evaluation

Let's put together the various pieces presented so far to see what evaluating a non-traditional type of program might look like. I will examine each of the three non-traditional programs presented earlier to see what evaluation methods will supply satisfactory data for the decisions that teachers in the program must make. This process, in essence, represents a simplification and application of the twelve-step evaluation model presented earlier by William P. Gorth.

Dietetic Technology Program

1. What should the content of the program include?

A needs analysis must be conducted to identify the proper course objectives. This needs analysis included a list of competencies by the primary controlling agency (the American Dietetic Association), a council of members from field agencies where students did their "practical"

work, and the experience of the instructor. The evaluation of this step would include the maintenance of current competency lists by the ADA, yearly reviews by the advisory council, and by questionnaires to both graduates of the program and institutions where students were subsequently employed.

2. What is the best instructional strategy to use?

The program is based on behavioral objectives. The instructor made an initial determination of what instructional strategy to use based on program objectives. Objectives were divided into classwork (lecture and seminar), learning activity packages (LAPS) and field work (actual on-the-job experience in hospitals participating in the dietary program). The evaluation was two fold. Each objective included one or more related criterion-referenced test items. Application of those test items indicate the proportion of the students mastering the objective. The teacher alternates instructional strategies for objectives where a less than satisfactory percentage of students achieved the objective. The teacher can, over time, develop the best instructional strategy(s) for each objective. Also, a questionnaire will need to be used to elicit appropriate data from field experiences. On-site observation and discussions with both students and supervisors in the local institutions will round out the evaluation data-gathering methods to use.

3. Were the objectives of the program learned by students as a group?

The criterion-referenced tests mentioned in #2 above will also provide the data needed to answer this question. If a less than satisfactory percentage of students learned a program objective, some sort of immediate revision is necessary. Either the objective is important enough to "try again" on the part of the instructor, or, perhaps the objective can be dropped.

4. Did individual students master course objectives?

Again, the criterion-referenced data is the key source of information, supplemented by the questionnaires, observations and interviews for the field experiences. Indi-

vidual students not mastering course objectives can be singled out for remediation or elimination from the program.

5. Were the students adequately prepared to be Dietetic Technologists by the program?

The data gathering instruments here will overlap with the data gathering instruments of #1 above. A questionnaire for both program graduates and employers should be used. Selected interviews will also be useful.

Minority Construction Contractor Program

1. What should the content of the program include?

The content was chosen by the college. This is an acceptable way to start a program as long as adequate evaluation data is gathered to revise and modify the program over time so that it meets the needs of the students in the program. Questionnaires and interviews with users should be used while they are in the program and after they have finished. Questionnaires and interviews would also be useful with successful local contractors and firms hiring the program graduates.

2. What is the best instructional strategy to use?

The criterion-referenced tests, based on course objectives, will supply this data. Also, since the population is "non-traditional," involving the students in deciding instructional strategies will be useful.

3. Were the objectives of the program learned by students as a group?

The criterion-referenced test items will provide the data needed for the traditional courses. The elements for self-support, confidence building, and awareness will probably require some combination of simulation, discussion, questionnaires, and interviews to determine how well the group achieved the objectives. Also, the non-traditional courses should probably be flexible enough so that student satisfaction would be as good an indicator as any.

4. Did individual students master the course objectives?

The data collected for #3 above would also apply here.

5. Were the students adequately prepared for acting as construction contractors?

This program is different from most others in that many of the students are already acting as construction contractors. Therefore, they have a good idea of where their deficiencies lie. The evaluation instruments will be the same as indicated under #1, i.e., questionnaires, interviews, and observation. Also, a criterion-referenced mastery test that has been properly validated might prove useful.

Retiree Program (focus on pre-retiree portion only)

1. What should the content of the program include?

The college and members of the program determined the content. Questionnaires, observations, and interviews with former program members, after retirement, will provide useful data as well as input from sociologists, psychologists, and financial experts.

2. What is the best instructional strategy to use?

The courses are all non-credit, and the acquisition of cognitive knowledge is less critical than in other programs, particularly career-oriented programs. Therefore, use of criterion-referenced tests will only tell part of the story. Student comments, gathered through questionnaires, interviews, and unobtrusive measures should play a very strong role in modification of instructional strategies.

3. Were the objectives of the program learned by students as a group?

The criterion-referenced test items will provide some data. But, in a program such as this with many affective components, simulations and unobtrusive measures should also be used.

4. Did individual students master the program objectives?

The same collection strategies as were used in #3 apply here.

5. Were the students adequately prepared for retirement by the program?

The methods listed for #1 would apply here—questionnaires, interviews and observations. Also, unobtrusive measures should be used.

Summary

This presentation has been brief. The decisions examined are subsets of the total number of decisions to be made, and the evaluation methods discussed are by no means all inclusive. However, the purpose of this presentation is to look at means of evaluating non-traditional programs. The thesis presented is that evaluation methods, if properly utilized, already exist to evaluate non-traditional programs as well as traditional programs. If the twelve step model proposed by William P. Gorth is followed and the principal tenets and guidelines of Peter Schriber and Eva Baker are adhered to, non-traditional programs can be successfully evaluated using the modern, up-to-date evaluation methods presently available.

Conference Roster

Dr. D.F. Adkisson
President
Cleveland State Community College
Cleveland, TN 37311

Richard Allan
National Evaluation System
P.O. Box 226
Amherst, Massachusetts 01002

Fadia Alvic
Graduate Student-UTK
3500 Sutherland # C-128
Knoxville, TN 37919

Dr. James H. Amburgey
Academic Dean
Hiwassee College
Madisonville, TN 37354

Kathy Aycock
Graduate Student-UTK
Route 12 Old Tazewell Pike
Knoxville, TN 37918

Thomas Ayers
Assistant Director
Off Campus Credit Programs-UTK
451 Communications and
University Ext. Bldg.
Knoxville, TN 37916

Dr. William D. Barber
Assistant Professor Agronomy-UTK
347 Plant Science Building
Knoxville, TN 37916

Ted A. Beach
Graduate Student-UTK
729 Schenley Road
Knoxville, TN 37919

Eunice Bell
Instructor of Nursing
Columbia State Community College
Columbia, TN 38401

Mrs. Nancy Boyd
Assistant Professor Off. Careers
Cleveland State Community College
Cleveland, TN 37311

Dr. Nolen Bradley
Dean of Instruction
Volunteer State Community College
Gallatin, TN 37066

Dr. Charles W. Branch
President
Chattanooga State Technical
Community College
4501 Amnicola Highway
Chattanooga, TN 37406

Dr. Donald V. Brown
Associate Professor-UTK
Knoxville, TN 37916

Dr. William H. Calhoun
Professor and Acting Head-UTK
Psychology Department
8 Ayres Hall
Knoxville, TN 37916

Dr. Jack Campbell
President
Walters State Community College
Morristown, TN 37814

John Campbell
Graduate Student-UTK
2211 Edgewood
Knoxville, TN 37919

L.B. Cebik
Associate Professor Philosophy-UTK
815 McClung Tower
Knoxville, TN 37916

Mary Helen Cebik
Administrative Assistant
College of Home Economics-UTK
Room 112
Knoxville, TN 37916

Harlan Chapman
Director of Admissions-Registrar
Dalton Junior College
Dalton, GA

Steve Chapman
Graduate Student-UTK
3500 Sutherland E-305
Knoxville, TN 37919

Dr. Jack Clemmer
Division Chairman of Social Science
Columbia State Community College
Columbia, TN 38401

Dr. William H. Coffield
Associate Dean Graduate Studies
Professor & Head Department of
Continuing and Higher Education
University of Tennessee
Knoxville, TN 37916

Ms. Winnie Cook, Director
Resources for Student Learning
Southeastern Community College
P.O. Box 151
Whiteville, NC 28472

Hank Cooper
Graduate Student-UTK
Chattanooga State Technical
Community College
Chattanooga, TN 37406

Carolyn Crowder
Graduate Student-UTK
1416 Lonas Drive
Maryville, TN 37801

Beatrice Curry
Instructor of English
Columbia State Community College
Columbia, TN 38401

Ray Davidson
Instructor of Electronics
Columbia State Community College
Columbia, TN 38401

Dr. Mack P. Davis
Vice President
Academic Affairs
East Tennessee State University
Johnson City, TN 37601

S.M. Denton
813 Broadway@ Gill
Knoxville, TN 37917

Dr. Gail W. Disney
College of Home Economics
University of Tennessee-Knoxville
Knoxville, TN 37916

Sister Dominica
Academic Dean
Aquinas Junior College
4210 Harding Road
Nashville, TN 37205

Dr. Cuyler Dunbar
President
Roane State Community College
Harriman, TN 37748

Dr. Joe Dzikielowski
Assistant Professor of Health
Cleveland State Community College
Cleveland, TN 37311

Dana Foster
Admissions Counselor-UTK
320 Student Services Bldg.
Knoxville, TN 37916

Dr. Liston M. Fox
Assistant Dean
College of Business Administration
University of Tennessee
52 Glocker Business Building
Knoxville, TN 37916

Miss Jo Fritchett
Assistant Professor of Office Careers
Cleveland State Community College
Cleveland, TN 37311

Dr. Warren E. Gaston
Assistant Dean of Admissions & Records
University of Tennessee
305 Student Services Bldg.
Knoxville, TN 37916

William Gorth
National Evaluation System
P.O. Box 226
Amherst, Massachusetts 01002

Dr. Donald Goss, Chairman
Division of Humanities
Volunteer State Community College
Gallatin, TN 37066

Dr. Leroy Graf, Professor and Head
History Department-UTK
1101 McClung Tower
Knoxville, TN 37916

Donald F. Hampton
Assistant Professor
Vocational-Technical Education
University of Tennessee
Knoxville, TN 37916

A.D. Haubenreich
Director of Readmissions
University of Tennessee
Knoxville, TN 37916

F. Haynes, Jr.
7026 Rockingham Drive
Knoxville, TN 37919

David L. Hazlewood
Leader Development Consultant
Holston Conference of the
United Methodist Church
Johnson City, TN 37601

Dr. James P. Hill, Jr.
Chairman of the Division of Natural
Sciences and Mathematics
Virginia Western Community College
Roanoke, VA

Bill Hodges
Counselor
Walters State Community College
Morristown, TN 37814

Everett Honaker
Assistant Director of Admissions
320 Student Services Building
University of Tennessee
Knoxville, TN 37916

L. Scott Honaker, Dean
College of Education
East Tennessee State University
Johnson City, TN 37601

Mr. Ed Howard
Assistant Professor of Economics
Cleveland State Community College
Cleveland, TN 37311

Wayne James
Graduate Student-UTK
3500 Sutherland #L-106
Knoxville, TN 37919

Joseph M. Keller
Dean of Instruction
Brevard Community College
Melbourne, Florida 32931

Ethyl Kintner
Graduate Student-UTK
Box K
Maryville, TN 37801

Milton M. Klein
Professor of History
McClung Tower
University of Tennessee
Knoxville, TN 37916

Dr. Quentin Lane
Dean of Instruction
Cleveland State Community College
Cleveland, TN 37311

Ms. Patricia Lebkuecher
Assistant Professor of History
Volunteer State Community College
Gallatin, TN 37066

Dr. James W. Loyd
Dean of Admissions & Records
East Tennessee State University
Johnson City, TN 37601

Fred H. Martin
Director of Institutional Research
Roane State Community College
Harriman, TN 37748

Joe McCormick
Instructor of Sociology
Columbia State Community College
Columbia, TN 38401

Dr. K. Owen McCullough
Assistant Professor
Department of Continuing &
Higher Education
University of Tennessee
Knoxville, TN 37916

Robert I. McElory
Dean of Student Services
Walters State Community College
Morristown, TN 37814

Dr. Arthur R. MacFadden
Associate Professor
Business and Commerce Division
Chattanooga State Technical
Community College
Chattanooga, TN 37406

Frank McKenzie
Graduate Student-UTK
Route 3
Madisonville, TN 37354

Celeste McNally
Graduate Student-UTK
Andy Holt Avenue
Knoxville, TN 37916

Don Mikula
Dean of Instruction
Shelby State Community College
Memphis, TN 38104

William A. Miller, Jr.
Assistant Dean of Engineering
University of Tennessee
121 Perkins Hall
Knoxville, TN 37916

Mrs. Shelby Millsaps
Instructor in Nursing
Cleveland State Community College
Cleveland, TN 37311

Anne P. Minter
Chairman, Math-Science Division
Roane State Community College
Harriman, TN 37748

Dr. Louis Moore
Division Chairman of Social Science
Columbia State Community College
Columbia, TN 38401

John Muldowny
Associate Professor
Department of History-UTK
1106 McClung Tower
Knoxville, TN 37916

Shirley Napier
Graduate Student-UTK
1611 Laurel Avenue #1005
Knoxville, TN 37916

Mrs. Martha Newman
Instruction Specialist
Cleveland State Community College
Cleveland, TN 37311

Alvin H. Nielsen, Dean
College of Liberal Arts-UTK
226 Ayres Hall
Knoxville, TN 37916

Mrs. Penelope Overend
Instructor of English
Cleveland State Community College
Cleveland, TN 37311

James C. Perry
Chairman, Division of Education
Roane State Community College
Harriman, TN 37748

Dr. John Peslak
Assistant Professor of Chemistry
Cleveland State Community College
Cleveland, TN 37311

Dr. Milton Phillips
State Board of Regents
1161 Murfreesboro Road
Nashville, TN 37217

Gilma N. Preus
Assistant to President
Shelby State Community College
Memphis, TN 38104

Paul K. Preus
Director Higher Education
Memphis State University
Memphis, TN 38152

Paul T. Pullen
Chairman-Humanities
Virginia Western Community College
Roanoke, VA 24015

Dr. John W. Prados
Vice President for Academic Affairs
University of Tennessee
705 Andy Holt Tower
Knoxville, TN 37916

Warree Rakiti
Graduate Student-UTK
U.T. Station
Knoxville, TN 37916

Dr. Earl M. Ramer
Professor
Department of Continuing &
Higher Education
University of Tennessee
Knoxville, TN 37916

Hal Ramer
President
Volunteer State Community College
Gallatin, TN 37066

Richard Rice
Assistant Professor-History
University of Tennessee
Knoxville, TN 37916

Dr. Paul A. Ritch
Division Head, Business & Commerce
Chattanooga State Technical
Community College
Chattanooga, TN 37406

George M. Roberts
Associate Director for Federal Affairs
Tennessee Higher Education Commission
908 Andrew Jackson State Office Bldg.
Nashville, TN 37219

Paul Sands
Dean of Instruction
Columbia State Community College
Columbia, TN 38401

Peter Schriber
National Evaluation System
P.O. Box 226
Amherst, Massachusetts 01002

Dr. Darrell D. Simmons
Dean of Instruction
Walters State Community College
Morristown, TN 37814

Lyle Smith
Assistant Professor of Psychology
Dyersburg State Community College
Dyersburg, TN 38024

Thomas F. Stovall
Associate Director Academic Affairs
Tennessee Higher Education
Commission
908 Andrew Jackson State Office
Building
Nashville, TN 37219

Sister Henry Suso, President
Aquinas Junior College
4210 Harding Road
Nashville, TN 37205

Gayle D. Thornton
Graduate Student-UTK
3101 Rifle Range Road
Knoxville, TN 37918

Joan Wagner
School of Nursing
University of Tennessee-Knoxville
Knoxville, TN 37916

Dr. David Watts
Assistant Professor of Biology
Cleveland State Community College
Cleveland, TN 37311

Trudy B. White
Graduate Student-UTK
3700 Sutherland Avenue #1-8
Knoxville, TN 37919

F.E. Wright, President
Jackson State Community College
Jackson, TN 38301

UNIVERSITY OF CALIFORNIA

ERIC® CLEARINGHOUSE FOR
JUNIOR COLLEGES

96 POWELL LIBRARY BUILDING
LOS ANGELES, CALIFORNIA 90024

EE 36

59

59
AUG 8 1980